Summer Program FGV/EMAp 2019

INTRODUCTION TO MACHINE LEARNING WITH PYTHON

Prof. Luis Gustavo Nonato

University of São Paulo - São Carlos - SP

Course Structure and Content

Course Structure and Content

- Principal Component Analysis

Course Structure and Content

- Principal Component Analysis
- Regression

Course Structure and Content

- Principal Component Analysis
- Regression
- Clustering

Course Structure and Content

- Principal Component Analysis
- Regression
- Clustering
- Classification

Course Structure and Content

- Principal Component Analysis
- Regression
- Clustering
- Classification
- Tree-based Regression and Classification

Course Structure and Content

- Principal Component Analysis
- Regression
- Clustering
- Classification
- Tree-based Regression and Classificatior

Data Science Central

ANALYTICS | BIG DATA | HADOOP | DATA PLUMBING | DATAVIZ | JOBS

The Algorithms Every Data Scientist Shoul Know

Class Dynamics

Our lessons will be divided in two parts:

Class Dynamics

Our lessons will be divided in two parts:

1. Theoretical content

Class Dynamics

Our lessons will be divided in two parts:

1. Theoretical content
2. Lab: Practice with Python

Class Dynamics

Our lessons will be divided in two parts:

1. Theoretical content
2. Lab: Practice with Python

Put ML algorithms to work in practice, using real data sets !!

Class Dynamics

Our lessons will be divided in two parts:

1. Theoretical content
2. Lab: Practice with Python

Put ML algorithms to work in practice, using real data sets !!

No deep knowledge in Python will be required.

Class Dynamics

Our lessons will be divided in two parts:

1. Theoretical content
2. Lab: Practice with Python

Put ML algorithms to work in practice, using real data sets !!

No deep knowledge in Python will be required.

The code will be as simple as possible and easily understandable, even for students not so experienced in computer programming.

Machine Learning: What and Why?

Machine Learning: What and Why?

"Given an email, is it a spam?"

Machine Learning: What and Why?

"Given an email, is it a spam?"

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

Machine Learning: What and Why?

"Given an email, is it a spam?"

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

Data can make up for the lack of knowledge !!

Machine Learning: What and Why?

"Given an email, is it a spam?"

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

Data can make up for the lack of knowledge !!

It is assumed that there is a "hidden" model/process capable of analyzing the content of an email and then decides whether it is a spam or not.

Machine Learning: What and Why?

"Given an email, is it a spam?"

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

Data can make up for the lack of knowledge !!

It is assumed that there is a "hidden" model/process capable of analyzing the content of an email and then decides whether it is a spam or not.

The goal is to *learn* the hidden model/process from data !!

Taxonomy

There are a variety of learning methods.

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1. **Predictive or Supervised**

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1. **Predictive or Supervised**
   - Regression: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1. **Predictive or Supervised**
   - Regression: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$
   - Classification: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1. **Predictive or Supervised**
   - Regression: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$
   - Classification: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   Given $\mathbf{x}$ predict y

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1. **Predictive or Supervised**
   - Regression: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$
   - Classification: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   Given $\mathbf{x}$ predict y
2. **Descriptive or Unsupervised**

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1. **Predictive or Supervised**
   - Regression: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$
   - Classification: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   Given $\mathbf{x}$ predict y
2. **Descriptive or Unsupervised**
   - Clustering

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1. **Predictive or Supervised**
   - Regression: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$
   - Classification: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   Given $\mathbf{x}$ predict y
2. **Descriptive or Unsupervised**
   - Clustering
   - Manifold Learning

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1. **Predictive or Supervised**
   - Regression: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$
   - Classification: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   Given $\mathbf{x}$ predict y
2. **Descriptive or Unsupervised**
   - Clustering
   - Manifold Learning
   - Relationship between instances

Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

1. **Predictive or Supervised**
   - Regression: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$
   - Classification: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$, $y \in \{c_1, \ldots, c_k\}$

   Given $\mathbf{x}$ predict y
2. **Descriptive or Unsupervised**
   - Clustering
   - Manifold Learning
   - Relationship between instances

The taxonomy above is not comprehensive, there are methods that do not properly fit in any of those two categories. For example, semi-supervised and reinforcement learning methods.

Recap

Summary of the Lecture:

Recap

Summary of the Lecture:

- This course covers important concepts and machine learning techniques

Recap

Summary of the Lecture:

- This course covers important concepts and machine learning techniques
- There are, though, relevant not coved topics (Neural Networks for instance)

Recap

Summary of the Lecture:

- This course covers important concepts and machine learning techniques
- There are, though, relevant not coved topics (Neural Networks for instance)
- We will adopt a very practical approach, with real data and applications.